SCIENCE AND TECHNOLOGY TEXT MINING: BASIC CONCEPTS

BY

Dr. Paul Losiewicz, Air Force Research Lab, Rome, NY.

Dr. Douglas W. Oard, College of Information Studies and Institute for Advanced Computer Studies, University Of Maryland, College Park, MD

Dr. Ronald N. Kostoff, Office of Naval Research, Arlington, VA.

(The views expressed in this paper are solely those of the authors, and do not represent the views of the Department of the Navy, Air Force Research Laboratory, or the University of Maryland)

ABSTRACT

This survey reviews a broad array of techniques that are becoming available to mine textual data. It presents initially a three function (data collection, data warehousing, data exploitation) text mining architecture consisting of a six step text mining process (source selection, text retrieval, information extraction, data storage, data mining, presentation). It then presents some of the most widely used data and text mining techniques, including clustering and classification methods (nearest neighbor, relational learning models, genetic algorithms) and dependency models (graph-theoretic link analysis, linear regression and decision trees, nonlinear regression and neural networks).

The survey finally illustrates some of their potential by describing the Office of Naval Research text mining pilot program. In the first year of that program, existing metadata from commercial bibliographic databases was used. There is presently an unacceptably long delay between the development of key component technologies for textual data mining and the deployment of the integrated tools that S&T sponsors need. The first year of the ONR text mining pilot program represents an initial attempt to bridge that gap. Important lessons have been learned about the use of text mining for management of science and technology research, but much remains to be done.

KEYWORDS: Data mining; text mining; information retrieval; OLAP; knowledge discovery in databases; data warehousing; information extraction; feature extraction; orthographic analysis; semantic analysis; statistical analysis; syntactic analysis; text retrieval; nearest neighbor; clustering; relational learning; genetic algorithms; link analysis; linear regression; decision trees; nonlinear regression; neural networks.

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1. REPORT DATE (DD-MM-YYYY)
14-07-2003

2. REPORT TYPE
Technical

3. DATES COVERED (FROM - TO)
xx-xx-1999 to xx-xx-2003

4. TITLE AND SUBTITLE
SCIENCE AND TECHNOLOGY TEXT MINING
BASIC CONCEPTS
Unclassified

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Losiewicz, Paul ;
Oard, Douglas W ;
Kostoff, Ronald N ;

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME AND ADDRESS
Office of Naval Research
800 N. Quincy St.
Arlington, VA22217

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME AND ADDRESS
Office of Naval Research
800 N. Quincy St.
Arlington, VA22217

10. SPONSOR/MONITOR'S ACRONYM(S)
ONR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
APUBLIC RELEASE
,

13. SUPPLEMENTARY NOTES

14. ABSTRACT
This survey reviews a broad array of techniques that are becoming available to mine textual data. It presents initially a three function (data collection, data warehousing, data exploitation) text mining architecture consisting of a six step text mining process (source selection, text retrieval, information extraction, data storage, data mining, presentation). It then presents some of the most widely used data and text mining techniques, including clustering and classification methods (nearest neighbor, relational learning models, genetic algorithms) and dependency models (graph-theoretic link analysis, linear regression and decision trees, nonlinear regression and neural networks). The survey finally illustrates some of their potential by describing the Office of Naval Research text mining pilot program. In the first year of that program, existing metadata from commercial bibliographic databases was used. There is presently an unacceptably long delay between the development of key component technologies for textual data mining and the deployment of the integrated tools that S&T sponsors need. The first year of the ONR text mining pilot program represents an initial attempt to bridge that gap. Important lessons have been learned about the use of text mining for management of science and technology research, but much remains to be done.

15. SUBJECT TERMS
Data mining; text mining; information retrieval; OLAP; knowledge discovery in databases; data warehousing; information extraction; feature extraction; orthographic analysis; semantic analysis; statistical analysis; syntactic analysis; text retrieval; nearest neighbor; clustering; relational learning; genetic algorithms; link analysis; linear regression; decision trees; nonlinear regression; neural networks

16. SECURITY CLASSIFICATION OF:

a. REPORT	b. ABSTRACT	c. THIS PAGE
Unclassified	Unclassified	Unclassified

17. LIMITATION OF ABSTRACT
Same as Report (SAR)

18. NUMBER OF PAGES
27

19. NAME OF RESPONSIBLE PERSON
Kostoff, Ronald
kostofr@onr.navy.mil

19b. TELEPHONE NUMBER
International Area Code
Area Code Telephone Number
703696-4198
DSN
-

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39.18

1. Introduction

Research management involves the effective and efficient acquisition, allocation, and disbursement of resources to advance the state-of-the-art of knowledge, and transitioning and tracking the research products as well. A critical component in the efficient management of research is maintaining awareness of the vast literature describing past, present, and future activities. Rapid growth of already enormous collections of myriad electronic texts is outpacing tools and especially skills for dealing with them.

Information retrieval systems that allow users to search large text collections using explicit word-based and phrase-based queries are now widely deployed, and in the hands of a skilled user they can be used to gain access to information from this broad array of sources. But traditional information retrieval systems treat documents as individual entities, placing the burden of recognizing relationships within and across documents on the user. Database systems began in a similar way, but the recent emergence of effective data mining techniques has expanded the range of ways in which users can interact with the data. Data mining essentially provides pattern-based retrieval, in which a pattern in the data is first discovered, and then that pattern is used to present information (the pattern itself or outlier data, perhaps) to the user. Online Analysis and Processing (OLAP) tools that can support interactive visualization of data mining results are becoming more common, and increasingly sophisticated visualization tools are becoming available as well.

Many large text collections include extensive metadata -- data about the texts in the collection. Existing data mining techniques offer a rich array of tools to exploit this data, and the recent development of practical (although imperfect) automatic techniques for extracting this sort of metadata from electronic texts offers a revolutionary potential for employing Textual Data Mining (TDM) techniques on a massive scale. The purpose of this paper is to survey the techniques that are available to support this task, and to relate them to the particular requirements of the management of science and technology.

Specific issues that arise repeatedly in the conduct of research management, and that could potentially be addressed by TDM, include:

- What Science & Technology is being done globally;
- Who is doing it;
- What is the level of effort;
- Where is it being done;
- What is not being done;

- What are the major thrust areas;
- What are the relationships among major thrust areas;
- What are the relationships between major thrust areas and supporting areas;
- What are the promising directions for new research;
- What are the innovations and discoveries?

The next section contains a TDM architecture that unifies information retrieval from text collections, information extraction from individual texts, knowledge discovery in databases, knowledge management in organizations, and visualization of data and information. At the core of this architecture is a broad view of data mining - the process of discovering nuggets and patterns in large collections of data - that is described in detail in Section 3. Section 4 illustrates how these ideas can be applied in practice, drawing upon examples from the recently completed first phase of the TDM program at the Office of Naval Research. The paper concludes by identifying some research directions that offer significant potential for improving the utility of TDM technology for research management.

2. Textual Data Mining Architecture

The term *Data Mining* generally refers to a process by which accurate and previously unknown information can be extracted from large volumes of data in a form that can be understood, acted upon, and used for improving decision processes [Apte, 1997]. Data Mining is most often associated with the broader process of *Knowledge Discovery in Databases (KDD)*, "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" [Fayyad et al., 1996]. By analogy, this paper defines *Textual Data Mining* as the process of acquiring valid, potentially useful and ultimately understandable knowledge from large text collections.

The distinction between text and data is an important one. *Data* are numeric or categorical values that acquire their meaning from a fully specified model (or *schema*) that relates the data to physical reality. *Text*, by contrast, is expressed in natural language, a format that can (at least in principle) convey any meaning. This paper addresses written texts that are stored as sequences of character codes, but similar approaches could be developed for document images or spoken language (see [Doermann, 1998] and [Foote, 1999] for summaries of some initial work on those modalities).

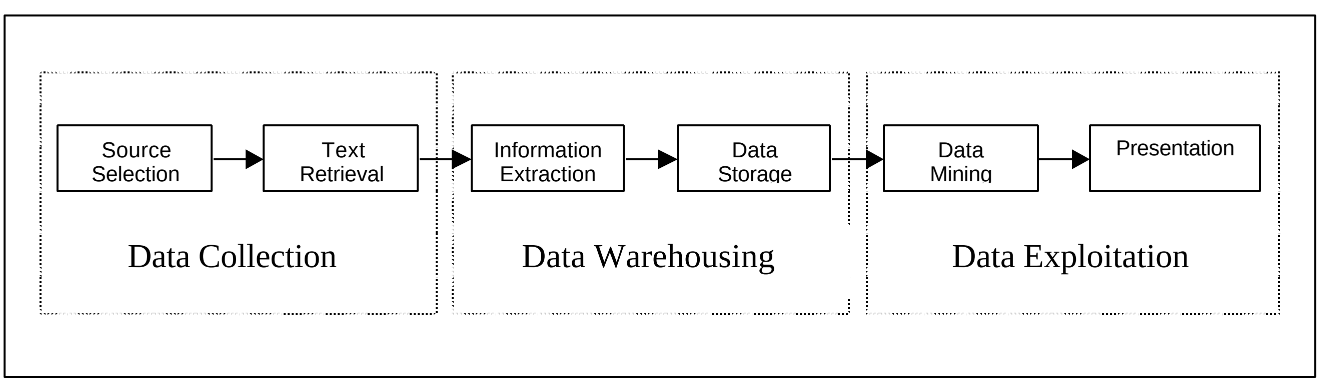
The distinction between data and text has important consequences for the formulation of TDM. Fayyad, et al. were interested in a KDD process that was *nontrivial*, meaning that they wished to find an expression in some language describing some subset of the data rather than simply retrieving that subset of the data as records in their original form. Texts are themselves linguistic expressions, and if the desired information were contained in a single text it would be entirely reasonable to simply find that text. The requirement that the patterns be *novel* can be omitted from the definition of TDM for the same reason.

Fayyad, et al. described a simplified five-step process KDD model that provided some degree of structure to the analysis of a broad range of KDD approaches. That process model included data selection, data preprocessing, data transformation, data mining, and data interpretation/evaluation. Figure 1 presents a similarly broad model for a six-step TDM process in which three main functions are performed:

- *Data Collection*, including Source Selection and Text Selection, which corresponds to Fayyad, et al.'s data selection step.

- *Data Warehousing*, including Metadata Assignment and Data Storage, which corresponds to Fayyad, et al.'s preprocessing step.

- *Data Exploitation*, including Data Mining and Data Presentation, which corresponds to



Fayyad, et al.'s last three steps.

Source Selection is the process of selecting sources to exploit. Source selection requires awareness of the available sources, domain knowledge, and an understanding of the goals and objectives of the data mining effort.

Text Selection is the process of discovering, selecting, and obtaining individual texts from the selected sources. This process might be fully automated or it may be performed interactively by a domain expert.

Metadata Assignment is the process of associating specific data with individual texts that is

Figure 1. The textual data mining architecture.

Data storage is the process of providing storage of and access to data. Data models that specify known relationships in the data can be stored with the data to facilitate subsequent processing.

Data Mining is the process of fitting models to data. The *data transformation* function that Fayyad, et al. identify is subsumed in this step for presentation purposes.

Presentation is the process of explaining and visualizing data mining results to support evaluation of data quality, assessment of whether the selected model is appropriate, and interpretation of the model.

Each step (except, perhaps, data storage) offers significant potential for user interaction, and some (e.g., source selection) might be done entirely manually. The remainder of this section gives an overview of the fundamental issues in each step of the TDM architecture.

2.1. Source Selection

There are myriad text sources that could be of use for textual data mining. Source selection has traditionally been performed manually based on a number of criteria, such as:

- **Coverage.** In order for TDM to be useful, the collection must contain texts that provide the needed information. Many TDM applications are subject-specific, so topical coverage is typically a key source selection criterion. Other criteria that may be important in specific applications are temporal coverage (e.g., earliest date, recency) and geographic coverage (e.g., place of publication).

- **Availability**. The rapid expansion of the Internet has made it practical to assemble collections that would previously have been impractical. For example, Los Alamos National Laboratories maintains freely accessible pre-print archives in Physics, Nonlinear Sciences, and Computer Science. At the same time, ubiquitous networking is simplifying the delivery of materials from commercially available text collections such as those provided by Dialog and Lexis/ Nexis. Access to legacy documents that exist only in hardcopy form is somewhat more problematic, since the cost of retrospective conversion to an appropriate digital format can be prohibitive in many applications.

- **Cost**. Although the Internet offers access to a wealth of freely available information, commercial interests often own the copyright on the specialized academic and trade journal texts that can be particularly useful for TDM. Many of these specialized text collections are already available in electronic form at special libraries and information centers. Site licensing arrangements that allow desktop access anywhere in an organization are becoming increasingly common, using either subscription or per-access pricing models.

- **Format**. Until recently the most widely available text collections included only abstracts, but the availability of full text is now becoming common as documents are captured in electronic form at the time they are generated. The brevity and uniform length of typical abstracts facilitate information extraction by limiting the potential for ambiguity, but full text collections provide far greater scope for the application of advanced information extraction algorithms. Some full text collections also contain potentially useful structural markup that can help guide information extraction by unambiguously indicating the extent of textual units such as paragraphs, tables, and lists. Common examples of structural markup include Standard Generalized Markup Language (SGML) and derivative formats such as Hypertext Markup Language (HTML) and Extensible Markup Language (XML),

- **Metadata**. Some text collections, most often those available for a fee from commercial sources, include extensive metadata that can be used directly in TDM applications. For example, the journal articles indexed in Science Citation Index include unambiguous citation links that can be used directly for co-citation analysis, and author fields constructed using name authority control can be used directly for co-authorship analysis. Similarly, use of a controlled vocabulary for subject indexing can facilitate topical clustering.

2.2. Feature Extraction

The next two steps, text selection and metadata assignment, can include automatic processing that exploits features extracted from each document. There are five fundamental sources of evidence about meaning in texts: orthographic, semantic, statistical, syntactic and usage analysis, and multiple sources can be combined to produce stronger evidence than could be obtained from a single source.

- **Orthographic Analysis.** The most basic form of analysis is the recognition of units of meaning in texts that are made up of character codes. In English, it is common to separate the text at white space (spaces, tabs, line breaks, etc.) and to then treat the resulting tokens as words. For languages that lack reliable orthographic evidence of word boundaries, one common approach is to use a sliding window to form overlapping *n*-character sequences that are known as "character *n*-grams" or "*n*-graphs".

- **Semantic Analysis.** The goal of semantic analysis is to relate the surface form of a text to a representation of the meaning that it expresses. Different words might be used to represent quite similar concepts, and morphological analysis can help deal with that by conflating several variants into a single root form. In some languages (e.g., English), a similar effect can be achieved through automatic suffix removal, a technique known as stemming. Another common semantic analysis approach is to exploit relationships among words (e.g., broader and narrower terms) that are encoded in a lexical knowledge structure such as a domain-specific thesaurus or a more general ontology such as WordNet. Simple knowledge structures such as term lists can also be used to recognize multi-word expressions such as idiomatic expressions that are not well handled by compositional techniques.

- **Statistical Analysis.** Statistical analysis of term usage frequency has proven to be quite useful as a source of evidence about meaning. The simplest approaches are based on counting the number of occurrences of each term (word, word root, word stem, *n*-gram, phrase, etc.). In collections with broad subject coverage, for example, it is common to find that the same phrase used in different contexts will represent different concepts. Statistical analysis of phrase co-occurrence can help to resolve this word sense ambiguity.

- **Syntactic Analysis.** Parsers with sufficient coverage and speed to process moderately large text collections are becoming available, and their output can provide several types of evidence about meaning. Part-of-speech analysis, for example, can help to disambiguate word sense, and syntactically identified phrases can provide a basis for further statistical analysis. Syntactic analysis can also be combined with semantic and lexical analysis to identify appropriate word boundaries in un-segmented languages such as Chinese and freely compounding languages such as German.

- **Usage Analysis.** The way in which a document is used can provide valuable cues about the document itself. For example, if the vast majority of the links to a document on the World Wide Web are from pages that mention known chemical compounds, it would be reasonable to infer that the document might be of interest to chemists. Oard and Kim (1998) identified four categories of potentially observable user behavior that could provide a basis for usage analysis: examination, retention, reference, and assessment.

2.3. Text Selection

Manual assignment of metadata is expensive, and automatic techniques for assigning metadata are, at least at present, computationally demanding. It can thus be beneficial to pre-select texts that are likely to be of value as an efficiency measure when dealing with large text collections. Some automated assistance with this process is typically needed, but integrating human judgement is also important because broad-coverage techniques that are optimized for automated processing of very large collections can not yet approach human performance on this task. Systems designed for this purpose are commonly referred to as *information retrieval* or, more specifically, *text retrieval* systems.

Modern text retrieval systems principally rely on orthographic, semantic, and statistical analysis. The usual approach is to use white space to identify word boundaries, followed by stemming to conflate words with similar surface forms into a common term. A weight is then computed for each term in every document using the frequency of the term in the document, the selectivity of the term (the fraction of the collection in which the term is found), and the length of the document. In *vector space* text retrieval, queries are represented in a manner similar to the documents, and the similarity of each document in the collection to the query is then computed as the normalized inner product of the document and query term weight vectors. In *probabilistic* text retrieval, a term weight is treated as the probability of relevance of a document to a query, conditioned on the presence of that term in the query. Probabilistic and vector space techniques are often combined with *Boolean* text retrieval, in which the presence or absence of a term or combination of terms can be explicitly required in the query specification. The principal advantage of vector space and probabilistic text retrieval over a purely Boolean approach is that lists of documents that are ranked in order of decreasing probability of relevance (or decreasing similarity to the query) allow users to interactively decide how many documents are worth examining. Unranked Boolean techniques, on the other hand, might be preferred when no user interaction is possible before the next processing stage. In either case, when the document collection is relatively stable it is common to preprocess the collection to produce an index structure on the feature set that can be searched in sub-linear time.

The utility of a text retrieval system depends strongly on how well the query is constructed, and that depends in turn on how well the user understands the collection and the way in which the indexed features can be used to select documents. It is usually fairly straightforward to find some topically relevant documents, but interactive inspection by the user is generally needed if the relevant documents must be more carefully separated from the irrelevant ones. An iterative query reformulation process such as Simulated Nucleation [Kostoff, 1997] can be used to speed this process, leveraging inspection of a few documents to produce a query that better separates relevant and irrelevant documents.

2.4. Metadata Assignment

The association of metadata with individual texts in large collections such as those provided by indexing and abstracting services and libraries has traditionally been a manual process, although interactive machine-aided indexing techniques are becoming more popular (cf., [Hlava et al., 1997]). TDM systems that seek to exploit multiple collections that contain manually assigned metadata face three challenges: differing content, differing formats, and differing vocabularies. Metadata content varies widely across collections, with each provider seeking to find a balance between the anticipated needs of their searchers and the expense of providing the metadata. For example, authors' names are often present in every collection, but organizations with which the authors are associated are less commonly provided. The recent effort to define a standard set of metadata that is known as the *Dublin Core* should simplify the design of TDM systems that exploit the standardized content, but metadata outside the Dublin Core framework will continue to require collection-specific processing. Some progress has already been made with standardizing metadata formats, but even widely accepted standards such as the Machine Readable Catalog (MARC) records used for library catalogs have national variants. The Resource Description Format (RDF) and the Extensible Markup Language (XML) together provide a standardized metadata format that is compatible with the Dublin Core, and TDM applications will benefit as these standards are more widely deployed.

Even when content and format are standardized, incompatible metadata values pose a challenge for TDM systems. *Authority control* is used to standardize the representation within a single system for person, place, organization, journal name, subject, and other items for which people might otherwise use different values to represent the same entity. There have been some large-scale efforts to merge the thesauri used to establish subject authority in different systems (c.f., [Selden and Humphries, 1996]), and fully automated techniques appear to offer some promise for this task [Gey, et al., 1999].

Many collections of gray literature lack reliable metadata, and the manual creation of metadata in such circumstances might be uneconomical. Automated information extraction algorithms offer a practical alternative in such cases. Information extraction can exploit deeper linguistic processing than the simple term frequency information used in text retrieval because information extraction can be applied to smaller and more coherent document sets. The usual approach is to exploit orthographic and syntactic analysis, typically using extraction rules that fire in two stages. The first step is to recognize the presence of a trigger word that selects a set of candidate syntactic patterns in which that word might occur. The syntactic context of the trigger word is then examined to identify the data to be extracted. These rules can be manually coded or they can be automatically learned from hand-labeled training texts (c.f., [Gallipili, 1996] or [Califf and Mooney, 1997]). Recent work on the application of unsupervised learning and statistical analysis to information extraction offers some promise that the requirement for hand-labeled training data can be relaxed somewhat [Riloff and Schmelzanbach, 1998].

Information extraction algorithms can be designed to operate without user interaction, making sharp decisions rather than producing ranked lists of candidates. Results from the Message Understanding Conferences (MUC) suggest that simple tasks such as date extraction and named entity recognition can be done quite accurately, but that more complex tasks in which relationships among data items must also be extracted are still somewhat error-prone. Lawrence, et al. have used information extraction techniques to extract bibliographic citations from gray literature on the World Wide Web [Lawrence et al., 1999].

The raw features exploited by text retrieval and information extraction algorithms offer a third source of metadata for TDM applications. For example, the similarity measure used in vector space text retrieval provides a natural basis for data mining techniques such as clustering.

2.5. Data Storage

The role of the data storage step is to accumulate the metadata, to record and manage information regarding the organization of that data, and to provide access to the metadata that is needed by the data mining step. Data mining applications often exhibit access patterns that differ markedly from those commonly experienced in other database applications. Existing relational and object oriented database systems may be adequate for relatively small-scale TDM applications, but careful attention to optimization may be required as the collection size grows. When multiple text collections are used in a TDM application, the metadata can be stored on a single machine or a federated database architecture in which collection-specific metadata are stored with each collection can be used. Mixing unstructured features such as vector representations of entire documents with more highly structured metadata such as subject terms drawn from a thesaurus is possible, but could pose additional implementation challenges.

2.6. Data Mining

The data mining step includes model selection, transformation of the data into a format appropriate for the selected model, and choice of a method for finding the appropriate parameters for that model. At a very general level, data mining models can be grouped by what they seek to produce (their *functions*). Functions traditionally associated with data mining include clustering models (producing categories), classification models (producing assignments to predefined categories), and dependency models (producing relationships between data elements). Viewed abstractly, a model is simply a parameterized set of ways in which the data can be used to produce a result, a criterion that describes which results are preferred, and an algorithm for searching the space of possible combinations. The output of the data mining step is the model and the parameters that have been found.

As an illustrative example, consider a TDM system that is designed to identify topically related news stories. A clustering model would be well suited to this task, since clusters could be used to represent topical relationships. If the data storage system contains the frequency of each word appearing in the title, a simple data transformation could reduce this to a Boolean value reflecting the presence or absence of each word. The parameters of a clustering model are the number of categories and the category to which each text is assigned. One simple preference criterion is to assign two texts to the same category if their titles share at least half the words in the shorter title. In this case, a greedy agglomerative clustering algorithm that sequentially makes pair-wise comparisons and either forms or joins clusters will quickly discover the optimal parameters for the model. Better clustering models can (hopefully!) be designed for this task, but the key parts of the data mining step are illustrated by this example.

2.7. Presentation

The presentation step generally seeks to serve three broad goals: helping the user understand the results, helping the user assess whether the chosen model was appropriate, and helping the user assess whether the quality of the data is adequate to support the desired analysis. It is often useful to provide a range of display capabilities and then allow users to select those that best match their goals and preferences. Presentation involves two sub-processes: summarization and display. Summarization is the process of producing abstract representations that emphasize some features of the data and suppress others. For example, the number of news stories assigned to a topical category might be included in a summary, but the title of each story might be omitted. Display is the process of representing the summary in a form that is appropriate for human perception. Graphical displays are often used because they capitalize on human visual perceptual abilities such as focus and pattern recognition, but other sensory channels (e.g., auditory display) can be useful in some cases. Research on information visualization has produced some tools that can exploit both metadata and text similarity (cf., [Rohrer, et al., 1998]). In TDM applications, those two sources can be augmented with the model parameters produced in the data mining step to produce richer visualizations (cf, Chen et al, 1998).

3. Data Mining Techniques

The TDM architecture shown in Figure 1 follows the traditional KDD architecture closely, first obtaining a large collection of numeric and/or categorical data and then fitting models to that data. A wide range of data mining techniques has been developed for use with numeric or categorical data. This section surveys some of the most widely used techniques. The classification system used is an augmented compendium of those proposed by [Westphal and Blaxton, 1998], [Gilman, 1998] and [Fayyad et al., 1996]. Textual Data Mining employs variations of many of these methods, including clustering, classification, and dependency modeling.

3.1. Clustering and Classification Methods

Clustering seeks to identify a finite set of abstract categories that describe the data by determining natural affinities in the data set based upon a pre-defined distance or similarity measure. Clustering can employ categories of different types (e.g., a flat partition, a hierarchy of increasingly fine-grained partitions, or a set of possibly overlapping clusters). Clustering can proceed by agglomeration, where instances are initially merged to form small clusters and small clusters are merged to form larger ones; or by successive division of larger clusters into smaller ones. Some clustering algorithms produce explicit cluster descriptions; others produce only implicit descriptions. *Classification* methods, by contrast, are used to assign data to predefined categories. A variety of techniques are available (e.g., decision trees, naïve Bayesian classifiers, and nearest neighbor classifiers).

3.1.1. Nearest Neighbor Methods

Nearest Neighbor algorithms support clustering and classification by matching cases internally to each other or to an exemplar specified by a domain expert. A simple example of a nearest neighbor method would be as follows: given a set $\boldsymbol{X} = \{x_1\ x_2\ x_3 \ldots\ x_n\}$ of vectors composed of n features with binary values, for each pair (x_i , x_j), x_i ¹ x_j,, create a vector v_i of length n by comparing the values of each corresponding feature n_i *of* each pair (x_i , x_j), entering a 1 for each n_i feature with matching values match and 0 otherwise. Then sum the v_i values to compute the degree of match. Those pairs (x_i , x_j) with the largest result are the nearest neighbors. In more complex nearest neighbor methods, features can be weighted to reflect degree of importance. Domain expertise is needed to select salient features, compute weights for those features, and select a distance or similarity measure. Nearest neighbor approaches have been used for text classification [Cost , S. et al, 1993].

3.1.2. Relational Learning Models

Relational learning models are inductive logic programming applications. Their foundation is logic programming using Horn clauses, a restricted form of first-order predicate logic. Logic programming describes relations on objects using declarative subject-predicate representations and uses classical deductive logic to draw conclusions. Data mining has been conducted by using inductive logic programming to generate database queries with a predicate logic query syntax.

3.1.3. Genetic Algorithms

Genetic algorithms can be used both for classification and for discovery of decision rules. Named for their Darwinist methodology, genetic algorithms use processing that is analogous to DNA recombination. A population of "individuals," each representing a possible solution to a problem, is initially created at random or drawn randomly from a larger population. Pairs of individuals combine to produce “offspring” for the next generation, and mutation processes are used to randomly modify the genetic structure of some members of each new generation.

Genetic algorithms perform categorization using supervised learning, training with a set of data and then using the known correct answers to guide the evolution of the algorithm using techniques akin to natural selection. Genetic methods have advantages over neural networks, because they provide more insight into the decision process. Sheth [1994] describes an example of the use of a genetic algorithm for text processing.

3.2. Dependency Models

Like clustering and classification, dependency models can be divided into two classes based on the nature of the dependencies being modeled. *Sequential dependency modeling* seeks to analyze temporal information encoded in the data to detect patterns of variation over time. This can be useful, for example, for performing citation analysis. *Static dependency modeling*, by contrast, seeks to explore relationships without regard to temporal order.

Dependency models can also be viewed from the perspective of the nature of the models that are considered. In *structural dependency modeling*, the goal is to discover unknown dependencies that exist among the data. These affinities can be expressed by qualitative rules such as “A causes B,” or by quantitative summaries such as "80% of the records that contain A and B also contain item C.” *Quantitative dependency modeling*, by contrast, assumes a specific model and seeks only to estimate of the parameters of that dependency relationship. *Regression* is an example of a quantitative dependency modeling technique.

3.2.1. Graph-theoretic Link Analysis.

Many dependency methods employ graph-theoretic approaches, structuring data into representations of causal links or influence diagrams to support sequential or static dependency analysis. The specific formalism adopted and the link and node semantics can vary with the domain, so effective design of the graph can require domain knowledge. Graph-theoretic models are generally easily visualized as node and link dependency structures.

Bayesian nets are probabilistic graphs that represent causal events as nodes and evidential dependencies as links. The evidence for belief in a leaf node accrues though the system of evidence links, based on the prior probabilities of the root nodes. The most tractable Bayesian networks are those that are instantiated in a Directed Acyclic Graph (DAG). The edges of the graph correspond to the causal or correlational relations that hold between the declarative evidence instantiated in the nodes. Associated with each node is a value corresponding to a belief or a probability whose value lies within the interval [0,1]. A Bayesian net is evaluated by computing the conditional values of the network nodes based upon inputs applied to the root nodes that have no incoming edges. Formally, for every node *n*, there is associated evidence *e*. P(*e*) is either given by *independent prior* (root) probabilities, or derived via Bayes' Rule. Bayes' Rule provides a tractable method for modeling the strengths of one's beliefs that an *x* is a *y*, given one's knowledge that *x* is also a *z*, but the requirement to estimate prior probabilities could limit the utility of Bayesian networks in TDM applications. To paraphrase Charniak, the numbers have to come from somewhere, and generally that reduces to informed guesses on the part of subject matter experts.

Freeman [1997] describes an interesting case of knowledge discovery within a graph-theoretic social network visualization. By rotating the network image around its center with a Virtual Reality Markup Language (VRML) viewer, three-dimensional clustering could be explored. After rotation, a dependency was discovered that would not have been revealed without rotation.

3.2.2. Linear Regression and Decision Trees

Linear regression (or *correlation*) methods are used to determine the relationships between variables to support classification, association and clustering. Variations include univariate and multivariate regression.

One common use of linear regression is to support generation of a decision tree. Decision trees are typically induced using a recursive algorithm that exhaustively partitions the data starting from an initial state in which all training instances are in a single partition, represented by the root node, and progressively creates sub-partitions that are represented by internal or leaf nodes. At each non-terminal node, the algorithm branches on values of the attribute that best discriminates between the remaining cases [White A. P. et al, 1994]. Each node will correspond to a rule characterizing some explicit property of the data, so generation of a decision tree is a restricted form of rule induction in which the resulting rules are mutually exclusive and exhaustive. Decision tree induction is fairly straightforward, but the results will only be useful if the available features provide sufficient basis meaningful categorization. To reduce computational complexity, heuristics are often applied to the selection of linear properties that implicitly omit from consideration the vast majority of potential rules. "This approach may leave valuable rules undiscovered since decisions made early in the process will preclude some good rules from being discovered later" [Gilmore 1998]. Rule extraction from decision trees can be used in data mining to support hypothesis validation.

3.2.3. Nonlinear Regression and Neural Networks

Nonlinear Regression algorithms are used to support classification, association and clustering. Domains whose properties are not well suited to linear partitioning can sometimes exploit non-linear algorithms such as feed-forward neural nets, adaptive spline methods, and projection pursuit regression.

Neural networks determine implicit rules where the classes invoked are not defined classically (e.g., not conforming to the principle of excluded middle), or are poorly understood and hence are not amenable to linear classification [McCulloch and Pitts, 1988]. A neural network is crafted from layers of neural units with various network typologies (number of units, layers, connections, connection strengths, and information propagation methods). One objection to the use of neural networks is that “the results often depend on the individual who built the model” [Gilmore 1998]. This is because the model, the network topology and initial weights, may differ from one implementation to another for the same data. A "black box syndrome" is also apparent in neural nets because there is rarely semantic insight to be gained from an inspection of the final state of the network. Unsupervised learning methods require no feedback from a domain expert -- instead, the network is used to discover categories based on correlations within the data.

Unsupervised learning using variant of the K-means clustering algorithm has been used with text in self-organizing maps of the type developed by Kaski et al [1995]. The alternative is supervised (or reinforcement) learning , in which expert feedback is given as part of the training set to indicate whether a solution is correct or incorrect.

The next section describes the Office of Naval Research FY98 pilot TDM program, the capabilities of TDM tools required, the types of TDM tools used in the first phase of the program, the lessons learned from using these tools, and some suggested future directions for improving capabilities.

4. Lessons Learned From ONR Textual Data Mining Pilot Program

The Office of Naval Research (ONR) established a pilot TDM program in 1998. The purpose of the initiative was to make more efficient use of global science and technology information to support program officers and other Science and Technology (S&T) managers in their planning, selecting, managing, reviewing, and transitioning functions. The specific goal was to develop the capability to answer the generic questions identified in Section 1, and to provide a documented basis for addressing the larger question of what ONR should be doing differently.

The program's technical structure for Fiscal Year 1998 (FY98) was based on experience with research on Database Tomography and bibliometrics at ONR over the past decade (Kostoff, 1991, 1992, 1993, 1994, 1995, 1997, 1998, 1999a). This combined approach contained four basic building blocks:

1) Information retrieval from S&T databases based on orthographic and statistical analysis and interactive query refinement;

2) Metadata obtained from bibliographic databases and through information extraction based on statistical analysis;

3) Data mining based on statistical and usage analysis; and

4) Presentation as a manually produced report.

The main differences between the FY98 program and previous Database Tomography and bibliometrics efforts were twofold:

1) Previous efforts concentrated on development of the overall process, whereas the FY98 effort focused on exploring the role of technical domain experts to support all study phases; and

2) A wider variety of technical databases was examined, resulting in greater emphasis on source selection and the need to accommodate a wider range of metadata format and content. Specific sources included:

- Technical papers and reports (Science Citation Index for basic Research, Engineering Compendex for applied research and technology, and NTIS for government-sponsored technical reports)
- Project Narratives (RADIUS for government-sponsored projects; IR&D for industry projects funded as overhead on Department of Defense contracts)
- Patents.

The lessons learned in the FY98 program are summarized in this section. These lessons are expected to inform future TDM work at ONR.

4.1. Iterative Query Reformulation

An iterative query refinement approach known as Simulated Nucleation (Kostoff, 1997) was used for all studies. This approach is summarized as follows. An initial query is entered into the database, and a sample of the retrieved records is selected. With the assistance of a technical expert, the sample is divided into two categories: those records relevant to the topic of interest, and those records non-relevant to the topic. Computational linguistics (phrase frequency and proximity analyses) are performed to identify phrase patterns characteristic of each category. The phrase patterns characteristic of the relevant category are used to modify the query to expand the number of relevant records retrieved, and the phrase patterns characteristic of the non-relevant category are used to modify the query to reduce the number of non-relevant records retrieved. The modified query is then used to perform retrieval, and the process is repeated until convergence to a consistent set of retrieved documents is obtained.

In past studies with Simulated Nucleation, about three iterations were typically required to obtain convergence. The resulting queries ranged in size from a dozen terms to a couple of hundred, depending on the relationship between the objectives of the study and the contents of the database. For example, one of the studies focused on S&T for aircraft platforms. The query philosophy was to start with the term AIRCRAFT, then add terms to the query that would increase the number of relevant S&T papers (e.g., VSTOL, HELICOPTER, LANDING GEAR, FUSELAGE) or would eliminate papers not relevant to S&T (e.g., CROP SPRAYING, BUFFALO TRACKING). The resulting Simulated Nucleation process for Science Citation Index (SCI) required three iterations and produced 207 terms, while the process converged after a single iteration for the Engineering Compendex (EC) with 13 terms.

Because of the technology focus of the EC, most of the papers retrieved using AIRCRAFT, HELICOPTER, or similar query terms focused on the S&T of the Aircraft platform itself, and were aligned with the study goals. The research focus of the SCI differed, with many of the papers retrieved focused on the science that could be performed from an Aircraft platform, rather than the S&T of the Aircraft platform itself, and those papers were not aligned with the study goals. Therefore, no adjustments were required to the EC query, whereas many terms were added to the SCI query using Boolean negation to eliminate non-relevant papers.

The iterative query approach provided an increased ratio of relevant to non-relevant papers, thus facilitating the subsequent information extraction and data mining steps. A larger number of relevant records in the originally targeted discipline, and in sometimes quite disparate related disciplines, were retrieved through this iterative query refinement process. This capability to discover documents in related disciplines has shown high potential for generating innovation and discovery from disparate disciplines (Kostoff, 1999b).

The FY98 pilot program concluded that iterative query refinement was the most crucial component of the data mining process, providing the foundational documents that set the upper limit on quality and comprehensiveness of any subsequent processing. However, the Simulated Nucleation technique was time and labor intensive. It required the judgement of a technical domain expert on: 1) the relevance of hundreds or thousands of technical abstracts, and 2) the selection of the highest impact query modification terms from thousands of candidates. Some experiments were performed to ascertain whether the step of dividing the retrieved records into relevant and non-relevant records could be bypassed while still yielding acceptable results. These experiments showed that it was possible to identify the more obvious query modification terms by examining the phrase frequency and proximity results from all the retrieved records without categorizing them first. However, there were many candidate query terms that could not be identified from visual inspection alone, independent of context. As a general rule, categorization into relevant/ non-relevant was necessary for accurate and comprehensive query modification and development.

Present efforts are focused on developing decision aid algorithms to accelerate the term selection component of the query modification process. The latest algorithms sort the candidate query modification terms by frequency of occurrence, and by ratio of occurrence in the two categories, relevant and non-relevant. Human judgment is still required to select terms from the relevant category that will be sufficiently specific to retrieve additional relevant documents without bringing along many additional non-relevant documents as well. Algorithms based on the co-occurence of multiple terms are now being investigated in an effort to further automate this process.

4.2. Bibliometrics

Bibliometrics are statistics related to the production, distribution and usage of documents, rather than the content of those documents. In the FY98 pilot program, sorted frequency lists were generated for authors, journals, organizations, countries, cited authors, cited papers, and cited journals using preexisting metadata. Bibliometric analyses were then performed on the retrieved records, and comparisons were made among the diverse technical disciplines studied (Kostoff, 1999a). This analysis allowed the critical infrastructure in each technical discipline (key authors, journals, institutions and countries) to be identified. This can be useful for finding credible experts to participate in workshops or serve on review panels, and for planning itineraries to visit productive individuals and organizations. For assessment purposes, the bibliometrics allowed the productivity and impact of specific papers, authors and organizations to be estimated. In this way, the critical intellectual heritage of each technical discipline can be characterized by identifying the most highly cited authors, papers, and journals. For perspective, context, and normalization, it is important to compare bibliometrics across technical disciplines, so that anomalies in any one discipline can be spotted more easily, and so that universal trends can be identified.

4.3. Phrase Frequency Analysis

Single word, adjacent double word phrases, and adjacent triple word phrases were extracted from the abstracts of the retrieved papers, and their frequencies of occurrence in the text were computed. A taxonomy was generated (top-down based on experience, bottom-up based on natural groupings of high frequency multi-word technical phrases, or some hybrid of top-down and bottom-up), and the appropriate technical phrases and their associated occurrence frequencies were placed in the appropriate categories. The frequencies in each category were summed, thereby providing an estimate of the level of technical emphasis of that category. For example, in the Aircraft S&T study, a taxonomy consisting of categories such as Structures, Aeromechanics, and Flight Dynamics was defined using a hybrid top-down and bottom-up approach. The sum of the frequencies of the phrases assigned by the technical expert to each of these categories was then computed.

This proved to be a very useful approach for estimating levels of emphasis.[1] When coupled with information about the desired level of emphasis for selected categories, judgments of adequacy or deficiency could then be made. Either a requirements-driven methodology could be used to relate what is being done to what is needed, or an opportunity-driven methodology could be used to relate what is being done to what the state-of-the-art will allow.

For the specific areas studied, phrase frequency analyses of requirements/ guidance documents were performed to obtain requirements-driven quantitative estimates of the desired levels of emphasis, and the phrase frequency results from the S&T documents were then compared with the phrase frequency results from the requirements/ guidance documents. Opportunity-driven desired levels of emphasis were estimated based on the intuition and judgement of technical experts, and compared with the phrase frequency results from the S&T documents. The two comparisons were used together to arrive at overall judgements regarding adequacy or deficiency.

A deeper taxonomy will naturally lead to greater resolution among subcategories, and thereby to greater specificity in the judgments of adequacy and deficiency that can be made. For example, if the lowest level materials category in a taxonomy of ship subsystems is MATERIALS, then a gross judgement of adequacy or deficiency of technical emphasis in the very broad category of MATERIALS is all that can be made. This does not help guide decisions because of the lack of specificity. If, however, the lowest level materials category in the taxonomy includes subcategories such as WELDED TITANIUM ALLOYS, then judgements as to the adequacy or deficiency of technical emphasis in WELDED TITANIUM ALLOYS can be made. The more detailed the sub-category, the more useful the result from a programmatic viewpoint, and the greater are the numbers of adequacy or deficiency judgements that can be made. However, the

---

[1] "Emphasis" is used here rather than "effort" because reporting rather than funding is the basis for the analysis.

greater the number or level of sub-categories, the lower the frequencies of the phrases required for statistical significance of each sub-category, the greater is the amount of work required, and the more expensive and time consuming is the study. Thus, a tradeoff between study time and costs, and quality of results required, must be performed.

It was also found useful to apply phrase frequency analysis to different database content fields to gain different perspectives. The Keyword, Title and Abstract fields are used by their creators for different purposes, and the phrase frequency results can provide a different picture of the overall discipline studied based on which field was examined. For example, in the Aircraft study, a group of high frequency Keywords was concentrated on longevity and maintenance; this view of the Aircraft literature was not evident from the high frequency phrases from the Abstract field, where lower frequency phrases had to be examined to identify thrusts in this mature technology area.

The contents of the Keyword field reflect summary judgements of the main focus of the paper's contents by the author or indexer, and represent a higher level description of the contents than the actual words in the paper or abstract. Thus, one explanation for the difference between the conclusions from the high frequency Keywords and Abstract phrases is that the body of non-maintenance Abstract phrases, when considered in aggregate from a gestalt viewpoint, is perceived by the author or indexer as oriented towards maintenance or longevity. For example, the presence of the material category phrase CORROSION in the Abstract could be viewed by the indexer as indicative of a maintenance-focused paper, since many maintenance problems are due to the presence of corrosion.

However, while there may be a difference in high frequency phrases between the two data sources, there may be far less of a difference when both high and low frequency phrases are considered. Thus, a second possible explanation is that, in some technical areas in different databases, there is more diversity in descriptive language employed than in other technical areas. Rather than a few high frequency phrases to describe the technical area, many diverse low frequency phrases are used. This could result from the research encompassing a wider spectrum of topics, each of which is receiving less effort. It could also result from the absence of a recognized discipline, with its accepted associated language. This would reflect the use of a combination of terminology from a number of diverse fields to describe concepts in the technical area. Another explanation is that maintenance and longevity issues are receiving increased attention, and the authors/ indexers may be applying (consciously or subconsciously) this 'spin' to attract more reader interest.

As another example, the Abstract phrases from the Aircraft study contained heavy emphasis on laboratory and flight test phenomena, whereas there was a noticeable absence of any test facilities and testing phenomena in the Keyword field. Again, the indexers may view much of the testing as a means to an end, rather than the end itself, and their Keywords reflect the ultimate objectives or applications rather than detailed approaches for reaching these objectives. However, there was also emphasis on high performance in the Abstract phrases, a category conspicuously absent from the Keywords. The presence of descriptors from the mature technology or longevity categories in the Keywords, coupled with the absence of descriptors from the high performance category, provides a very different picture of the Aircraft research literature than does the presence of high performance descriptors and the lack of longevity and maintenance descriptors in the Abstract phrases.

This analytical procedure, and procedures based on phrase proximity that are described below, are not independent of the analyst's domain knowledge; they are, in fact, expert-centric. The computer-derived results help guide and structure the expert's analytical processes; the computer output provides a framework upon which an expert can construct a comprehensive story. The conclusions, however, will reflect the biases and limitations of the expert(s). Thus, a fully credible analysis requires not only expert domain knowledge on the part of the analyst(s), but also domain knowledge representing a diversity of backgrounds.

4.4. Phrase Proximity Analysis

After the high frequency phrases were identified using phrase frequency analysis, specific "theme phrases" of particular interest to the objectives of the study were selected. A statistical analysis was then performed to identify other phrases located in close proximity to each theme phrase, and those phrases most often associated with the theme phrase were identified. The process was applied both within and across fields to generate a variety of association results.

Applied to ‘infrastructure’ oriented database fields used for bibliometric analysis (title, author, journal, organization, and country), proximity analysis identified the key authors, journals and organizations closely related to specific technical areas of interest. This is particularly useful when attempting to define the infrastructure for an unfamiliar area. Applied to the Abstract field, proximity analysis allowed closely related themes to be identified. This may be of particular value in identifying low frequency phrases closely associated with high frequency themes; the so-called "needle-in-a-haystack." The background and perspective of the technical expert were extremely important in this process, since the core requirement is to recognize a "signal" from a substantial amount of “clutter”.

Applying proximity analysis to the Abstract field also allows taxonomies with relatively independent categories to be generated using a "bottom-up" approach. This is a potentially powerful capability because taxonomies are used in all phases of S&T performance and management, so a technique that can generate credible taxonomies semi-autonomously in relatively undeveloped disciplines would have many applications. Presently, the taxonomies are generated by: 1) selecting many high frequency themes using phrase frequency analysis; 2) identifying co-occurring phrases located near each selected theme phrase in the text; and 3) then grouping into categories related theme phrases that share more than a threshold number of co-occurring phrases. The process is somewhat labor intensive at present, but there is potential for substantial automation with attendant reductions in time and labor.

Applied across fields, proximity analysis allows access to complementary and disparate literatures that contain themes related to the target literature. This approach has a high potential for innovation and discovery from other disciplines (Kostoff, 1999b). Finally, proximity analysis has proven to be useful for estimating levels of emphasis for technical areas that are closely associated with specific technical sub-areas.

4.5. Technical Domain Expertise

The FY98 experience showed conclusively that high-quality data mining requires the close involvement of technical domain expert(s) in information retrieval, phrase frequency and proximity analyses, and presentation. Multiple perspectives are often needed to detect data anomalies -- both multiple domain experts with diverse backgrounds and data mining experts who have analyzed many different disciplines are needed.

Adopting a long-range strategic view, the main output from a data mining is technical experts who have had their horizons and perspectives broadened substantially through participation in the data mining process. The data mining tools, techniques and tangible products are of secondary importance relative to the expert with advanced capabilities who could be a long-term asset to an organization.

There was a steep learning curve required to integrate the domain expert with the computational tools that required substantial training time. The operational mechanics were not the problem; the major roadblock was the time required for the expert to understand how the tools should be applied to address the study's specific objectives, and how their products should be analyzed and interpreted. The problem stems from the fact that data mining requires additional skills beyond traditional science and engineering training and experience, and technical domain experts do not necessarily develop such skills in a traditional technical specialty career.

TDM cannot realize its full potential in S&T management if used only sporadically -- it must become an integral part of the S&T sponsor's business operations. Because of the learning curve, long-term involvement of experts with data mining experience in a particular topic area is desirable. A strategic plan that presents an S&T sponsor's TDM in this larger context is therefore needed to insure that data mining integration is implemented in a cost-effective manner. Such a plan should identify the different ways TDM would support aspects of the S&T sponsor's operations such as planning, reviews, assessments, and oversight response. Each of these applications has different objectives, metrics to address those objectives, and data requirements for each metric. The strategic plan should also address the role of TDM in the context of the organization's overall data mining effort, relationships with data mining efforts for similar technical areas by other organizations, and how the required types of technical expertise can be integrated most efficiently into the TDM process. Different types of experts and different suites of TDM tools may be required for different tasks.

A strategic plan allows a top-down approach to TDM in which the desired objectives are the starting point. This allows the sources and TDM tools required to satisfy the objectives to be identified, and planned for, in advance. Without such a plan, only existing sources and tools, generally those that were selected for other purposes, would be available. Such an approach would limit the results that could be obtained, forcing the use of whatever metrics the existing sources and tools will support, whether or not these metrics are most appropriate to satisfying the overall objectives of the application.

4.6. Cost and Time Estimates

The cost and time required for a TDM study will depend on the scope of the study and quality of the final product desired. For studies structured like those in the FY98 pilot program, the cost and time required will depend on:

- The complexity of the query, the number of query iterations, and the level of analysis effort applied to each iteration;
- The number of fields processed;
- The number of computational techniques employed and the number of different applications of each technique;
- The number and sophistication of the data mining techniques that are used; and
- The complexity of result presentation and interpretation.

A complete TDM study could range from a simple query in a focused technical field that could be performed by a program officer with little specialized training to a complex query refinement and analysis process that would require contractor support. The costs associated with these studies could range from zero out-of-pocket expense for a simple query to six figures for the generation of complex queries and sophisticated analyses and the time required for such studies could range from minutes to months.

4.7. Update

Subsequent to the 1998 ONR text mining pilot program, numerous science and technology text mining studies have been performed (Kostoff et al, 2000a, 2000b, 2001, 2002, 2003, 2004). These studies have evolved from a strong manual focus, especially in the clustering, to a semi-automated focus. Present text mining studies use both concept clustering (words/ phrases) and document clustering. The concept clustering uses factor matrix and multi-link hierarchical aggregation approaches, and the document clustering uses partitional and hierarchical aggregation approaches. In both cases, intensive manual labor and technical judgement are still required to extract information from the clusters and form the final taxonomy.

5. Conclusion

This survey has reviewed a broad array of techniques that are becoming available to mine textual data, and has illustrated some of their potential by describing the Office of Naval

Research TDM pilot program. In the first year of that program, existing metadata from commercial bibliographic databases was used. There is presently an unacceptably long delay between the development of key component technologies for textual data mining and the deployment of the integrated tools that S&T sponsors need. The first year of the ONR TDM pilot program represents an initial attempt to bridge that gap. Important lessons have been learned about the use of textual data mining for management of science and technology research, but much remains to be done.

Acknowledgements

The authors wish to express their appreciation to Robert Strange for his help with preparing the original survey from which this paper evolved, and to Surajit Chaudhuri, Marti Hearst, David Jensen, Dunja Mladenic, Vijay Raghavan, Dagobert Soergel, and the anonymous reviewers for their insightful comments on earlier versions of this paper. The work of Douglas Oard has been supported in part by DARPA contract N6600197C8540.

6. Bibliography

Apte, C. (1997). Data Mining An Industrial Research Perspective. *IEEE Computational Science and Engineering*. v 4.

Califf, M.E., and Mooney, R.J., (1997). *Applying ILP-based Techniques to Natural Language Information Extraction: An Experiment in Relational Learning,* Workshop Notes of the IJCAI-97 Workshop on Frontiers of Inductive Logic Programming, pp. 7-11, Nagoya, Japan, August.

Chen, H., Houston, A.L., Sewel, R. R., and Schatz, B.R., (1998). Internet Browsing and Searching: User Evaluations of Category Map and Concept Space Techniques, *Journal of the American Society for Information Science*, 49(7):582-603..

Cost, S., Salzberg, S. (1993). A Weighted Nearest Neighbor Algorithm for Learning with Symbolic Features. *Machine Learning*, 10, Page 57.

Doermann, D. (1998). The Indexing and Retrieval of Document Images: A Survey, *Computer Vision and Image Understanding*, June,.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine* 17(3): Fall, 37-54.

Foote, J. (1999). An Overview of Audio Information Retrieval, *ACM-Springer Multimedia Systems*, , to appear. Available at http://www.fxpal.xerox.com/people/foote/

Freeman, L. (1997). Using Available Graph Theoretic or Molecular Modeling Programs in Social Network Analysis [http://tarski.ss.uci.edu/new.html].

Gallippi, A. *Automatic Cross-Language Proper Name Determination in Text using Robust Methods*, Ph.D. Thesis, University of Southern California, Los Angeles, CA.

Gey, F., Chen, H-M., Norgard, B., Buckland, M., Kim, Y., Chen, A., Lam, B., Purat, J., Larson, R. (1999). Advanced Search Technologies for Unfamiliar Metadata, in *Third IEEE Meta-Data conference*, Bethesda, MD, April.. Available at http://www.sims.berkeley.edu/research/metadata/papers.html

Gilman, M. (1998). *Nuggets™ and Data Mining* Data Mining Technologies Inc White Paper, .

Hlava, M. M. K., Hainbebach, R., Belanogov, G., Kuznetsov, B. (1997). Cross-Language Retrieval - English/Russian/French. In S*ymposium on Cross-Language Text and Speech Retrieval*, Technical Report SS-97-05, American Association for Artificial Intelligence Available at http://www.glue.umd.edu/~dlrg/filter/sss/

Kostoff, R. N. (1991). Database Tomography: Multidisciplinary Research Thrusts from Co-Word Analysis. Proceedings: *Portland International Conference on Management of Engineering and Technology, October 27-31.*

Kostoff, R. N. (1992). Research Impact Assessment. *Proceedings: Third International Conference on Management of Technology.* Miami, FL. Larger text available from author.

Kostoff, R. N. (1993). Database Tomography for Technical Intelligence. *Competitive Intelligence Review.* 4:1.

Kostoff, R.N. (1994). Database Tomography: Origins and Applications. *Competitive Intelligence Review, Special Issue on Technology, 5:1.*

Kostoff, R.N. et al (1995) System and Method for Database Tomography. *U.S Patent Number 5440481.*

Kostoff, R.N., Eberhart, H.J. and Toothman, D.R. (1997). Database Tomography for Information Retrieval. *Journal of Information Science.* 23:4.

Kostoff, R.N., Eberhart H.J., and Toothman, D. R. (1998). Database Tomography for Technical Intelligence: A Roadmap of the Near-Earth Space Science and Technology Literature. *Information Processing and Management.* 34:1.

Kostoff, R. N., Eberhart, H. J., and Toothman, D. R. (1999a). Hypersonic and Supersonic Flow Roadmaps Using Bibliometrics and Database Tomography. *JASIS.* 15 April.

Kostoff, R.N., 1999b. Science and Technology Innovation. *Technovation,* 19:10, 1999.

Kostoff, R. N., Braun, T., Schubert, A., Toothman, D. R., and Humenik, J. (2000a). Fullerene Roadmaps Using Bibliometrics and Database Tomography. *Journal of Chemical Information and Computer Science.* 40:1. 19-39. Jan-Feb.

Kostoff, R. N., Green, K. A., Toothman, D. R., and Humenik, J. (2000b). Database Tomography Applied to an Aircraft Science and Technology Investment Strategy. *Journal of Aircraft*, 37:4. 727-730. July-August.

Kostoff, R. N., and DeMarco, R. A. (2001). Science and Technology Text Mining. *Analytical Chemistry.* 73:13. 370-378A. 1 July.

Kostoff, R. N., Del Rio, J. A., García, E. O., Ramírez, A. M., and Humenik, J. A. (2001). Citation Mining: Integrating Text Mining and Bibliometrics for Research User Profiling. *JASIST.* 52:13. 1148-1156. 52:13. November.

Kostoff, R. N., Tshiteya, R., Pfeil, K. M., and Humenik, J. A. (2002). Electrochemical Power Source Roadmaps using Bibliometrics and Database Tomography. *Journal of Power Sources.* 110:1. 163-176.

Kostoff, R. N., Shlesinger, M., and Malpohl, G. (2003). Fractals Roadmaps using Bibliometrics and Database Tomography. *Fractals*. December.

Kostoff, R. N., Shlesinger, M., and Tshiteya, R. (2004). Nonlinear Dynamics Roadmaps using Bibliometrics and Database Tomography. *International Journal of Bifurcation and Chaos*. January.

Lawrence, S., Giles, C.L., and Bollacker, K. (1999). Digital Libraries and Autonomous Citation Indexing, *Computer*, 32(6):67-71.

McCulloch, W.S. and Pitts. W. (1988). A logical calculus of ideas immінint in nervous activity, *Neurocomputing: Foundations of Research* (Anderson, J. A. and Rosenfeld, E., Eds). MIT Press, Cambridge MA.

Oard, D. W. and Kim, J. (1998). Implicit Feedback for Recommender Systems, *AAAI Workshop on Recommender Systems*, Madison, WI.. Available at http://www.glue.umd.edu/~oard/research.html.

Riloff, E., Schmelzenbach, M. (1998). An Empirical Approach to Conceptual Case Frame Acquisition. In *Proceedings of the Sixth Workshop on Very Large Corpora*, Montreal, August. Available at http://www.cs.utah.edu/~riloff/publications.html

Rohrer, R. M., Ebert, D. S., Sibert, J. L. (1998). The Shape of Shakespeare: Visualizing Text using Implicit Surfaces. In *Fourth IEEE Symposium on Information Visualization*, Durham, NC, October.

Selden, C. R. and Humphries, B. L. (1996). Unified Medical Language System, *Current Bibliographies in Medicine* 96-8, National Library of Medicine,. Available at http://www.nlm.nih.gov/pubs/cbm/umlscbm.html.

Sheth, B. **(**1994). A Learning Approach to Personalized Information Filtering, *Master's Thesis*, MIT.

Westphal, C., Blaxton T. (1998). *Data Mining Solutions*. John Wiley and Sons, New York, NY..

White A. P., Liu, W. Z. (1994). Bias in Information-Based Measures in Decision Tree Induction. *Machine Learning*, 15, 321-329.